WP. 15
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (ii): Editing administrative data and combined sources

EDITING STRATEGIES FOR VAT DATA

Supporting Paper

Prepared by Peter Kruiskamp, Statistics Netherlands[1]

I. INTRODUCTION

1. For Dutch enterprises, business surveys for Economical Statistics are considered a huge burden, and, especially for the small and medium-sized enterprises, reducing this administrative burden is a politically 'hot' item. It is therefore an important issue for Statistics Netherlands to reduce the use of sample surveys considerably, and to emphasise the use of business registers, such as Dutch tax registers.

2. Statistics Netherlands has started an ambitious project to redesign the processes of the Business-Statistics. Leading objectives in this project are a reduction of the response burden for small and medium-sized enterprises, and an increase of efficiency, while maintaining the current level of quality. The use of VAT (Value Added Tax)-turnover fits nicely within these objectives. Survey data collection by sampling will be reduced to a minimum, while the large amount of VAT-data could –in principle– ensure better quality of the Short-Time Statistics. There are a number of complications, however.

3. When VAT-data is converted into statistical data so it can be used for the Short-Term Statistics, there are a number of difficulties to overcome. This is discussed in section 2.

4. At the moment VAT data is already used in the Short-Term Statistics as an auxiliary variable. This, as well as the future situation, is discussed in more detail in section 3.

5. For proper use of VAT-data as data source for turnover data, it is important that an editing method is used that minimises the amount of erroneous data, without losing too much information. In section 4 an editing strategy is proposed, the emphasis in this paper will be on micro editing.

II. CONVERTING VAT-DECLARATIONS INTO STATISTICAL DATA

6. In the Netherlands, tax authorities collect VAT-data for obvious reasons. Enterprises themselves declare VAT within one month after the period under review. The Dutch tax authorities usually consider this VAT-declaration to be equal to the final VAT-assessment. However, when Dutch tax authorities find it necessary, a correction is made in the final VAT-assessment, usually a few months after the VAT-declaration was submitted. For the Short-Term Statistics, these corrections are not available in time to be

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

used. Therefore, only the original VAT-declarations are used. Depending on the size of the enterprise, VAT-declarations are submitted on a monthly, quarterly or yearly basis.

7. Dutch enterprises, or more specifically the underlying legal units, can divide themselves in fiscal units of their own choice (within certain restrictions) in order to declare tax. These fiscal units are not suitable for statistical purposes. Therefore, they have to be converted into statistical units. The composition of the fiscal units is known, and the so called fiscal persons can be linked to legal units, which in turn can be linked to statistical units, also called enterprises.

8. Complications arise since not every enterprise is required to declare VAT. Some activities are exempt from VAT. When the nett annual VAT stays below a certain level, no assessment is required. Furthermore, small enterprises submit their VAT declarations on a yearly basis; these are not useful for the Short-Term statistics.

9. Above-mentioned complications are usually (but not always) relatively small. A bigger problem is the loss of information on statistical units due to failure to link them unambiguously to fiscal units. Therefore, large enterprises cannot be monitored using VAT-declarations, and even for middle-sized enterprises this is a relatively big problem. Generally, for the small and middle-sized enterprises the amount of statistical units that cannot be unambiguously linked to fiscal units is about 20% (Groen *et al.*, 2007). Procedures are being developed to improve the conversion, which will not be discussed in this paper.

10. Another issue related to the transformation of the tax units into statistical units, is what to do with records that do not occupy a 100% ‘coverage percentage’ and / or ‘filling percentage’. The ‘coverage percentage’ is the percentage of legal units describing an enterprise that can be linked to fiscal units, required to submit VAT-declarations. The ‘filling percentage’ is the percentage of legal units describing an enterprise that can be linked to fiscal units, which actually submitted the VAT-declaration.

III. USAGE OF VAT-DATA

A. Current usage of VAT-data

11. Presently VAT-turnover is used as an auxiliary variable in order to improve sampled survey data. Specifically VAT-data is used for the following purposes:

- Checking the plausibility of survey-turnover with corresponding VAT information
- Correcting survey-turnover with donor-information from VAT-data
- Imputation of turnover data with VAT-data due to non-response
- Suppletion of turnover data with VAT-data due to the design of the survey (only for small enterprises)
- Weighing of the survey using VAT information

All these steps are important to obtain reliable results for the Short-Term Statistics. The present method of editing VAT-turnover is very rough (see section IV). In this situation this is justified however, since the VAT register contains a huge amount of data, and VAT-turnover is used only as an auxiliary variable.

B. Future usage of VAT-data

12. In the future situation, for only those business sectors where the use of VAT-data satisfies certain quality restrictions, enterprises with a number of working persons (WP) up to 49 will no longer be part of the survey. For these enterprises turnover will be estimated using VAT-data only. For larger enterprises this is not feasible, since there is too much risk of information loss due to complex transformations from fiscal units to statistical units, as mentioned before.

13. This new approach of estimating Short-Term Statistics for the small and medium-sized enterprises has important implications on the method of editing VAT-data. The quality of VAT-turnover data will have to be at the level of the present survey data. Therefore, a more sophisticated editing method will have to be implemented, in order to minimise the amount of erroneous turnover data while preserving as much VAT turnover information as possible.

IV. EDITING STRATEGY

14. Tax declarations have to be edited due to measurement and processing errors. Presently, during the process of converting rough VAT-declarations into turnover information on statistical units, VAT-data is edited at two separate stages. In the future situation, an approach using data editing techniques at three stages in the process is suggested (Hoogland and Van Haren, 2007). These three stages are discussed in the following sections.

A. Micro editing of VAT-data on the level of fiscal units

15. Presently, the VAT-data that is received from the Dutch tax authorities is scanned for huge outliers. When VAT-turnover of a fiscal unit is larger than € 100.000.000, and at least 10 times the mean value of the preceding periods in the same and previous year, it is considered a suspicious value. These suspicious records are then edited manually. Following a set of criteria, it is decided if the record is to be corrected, discarded or allowed to pass (Van Loo, 2007).

16. This very basic criterion for suspicious records detects only 20 to 30 extreme values a year. The idea behind this method is that mainly key- and scan-errors will be detected. Until 2004, most tax declarations were keyed in manually from written declarations. 2004 was a transition year, in which most declarations were scanned. From 2005 on it is required for (almost) all enterprises to submit their VAT-declarations electronically. Therefore huge errors as described will be rare nowadays, as is supported considering the number of outliers detected.

17. The editing procedure has not become completely obsolete, since extreme outliers are still detected. There are some drawbacks to the method, though. Using the mean value as reference data can result in the situation that outliers will not be detected. For instance, when previously an outlier is allowed to pass, this will increase the mean value for future periods considerably, causing even higher outliers not to be detected anymore. Therefore, it would be an improvement to use the median value. In order to account for seasonal and similar effects, the use of a time series model would be preferable to calculate a proper turnover estimation.

18. Furthermore, the procedure only detects extreme high values, while extreme low values should also be considered. The proposed criteria would then become:

$$O^t_{FU,i} > a \quad \text{and} \quad O^t_{FU,i} \geq c_1 \cdot O'^t_{FU,i} \qquad (1)$$

or

$$O'^{t}_{FU,i} > a \qquad \text{and} \qquad O^{t}_{FU,i} \leq \frac{1}{c_2} \cdot O'^{t}_{FU,i} \qquad (2)$$

$O^{t}_{FU,i}$ is the turnover and $O'^{t}_{FU,i}$ the reference value (mean or median turnover of the past two years, or an estimation using a time series model) of fiscal unit *i* for period *t*, *a* is the limiting turnover value and c_1 and c_2 the allowed factors. Note that in the present editing procedure only equation (1) is used, with *a* = 100.000.000 and $c_1 = 10$.

19. To determine *a*, c_1 and c_2 the final VAT-assessments could possibly be used, since these assessments can be considered to be a correction on the VAT-declarations, either by the Dutch tax authorities or by the enterprises themselves.

20. This editing method remains a very basic procedure due to lack of background information and insight in underlying mechanisms. On the level of fiscal units there is no information available of the WP of the fiscal unit, a variable necessary for stratification. Furthermore, the Standard Industrial Classification (SIC) used by the Dutch tax authorities is highly outdated and of little use for stratification and aggregation purposes. Moreover, the Dutch tax authorities keep their editing procedures on tax declarations secret to avoid misuse. For above reasons it is difficult to develop a more sophisticated editing method on the level of fiscal units. Nevertheless, Hoogland and Van Haren (2007) propose a method to refine the micro-editing procedure, using inter quartile distances to determine boundary values.

B. Data handling on the level of statistical units

21. In the Short-Term Statistics, all slightly suspicious VAT-data are discarded, in order to obtain a data set that contains as less erroneous turnover data as possible. VAT-declarations are discarded if no information is known for the reference data, which is the turnover of the concerning enterprise for (*t*-1) (*t* being a month or quarter, depending on the period of the Short-Term Statistics), or if VAT-turnover is zero or negative for either *t* or (*t*-1). Furthermore, declarations with a covering percentage or filling percentage less than 100% are considered incomplete and are not regarded. This adds up to a loss of data of about 25% of the 80% successfully linked enterprises.

22. For the remaining records the VAT-turnover is normalised with the median of the concerning publication cell (SIC x class of WP), and compared to the normalised value at (*t*–1). When the two values differ more than a factor 40, it is considered an outlier, and is also discarded. This occurs in approximately 0.4% of the cases. The value closest to the normalised VAT-turnover at (*t*–2) is considered the ‘true’ value (Van Velzen and Van de Pol, 2006).

23. As mentioned before, this all works fine when VAT-turnover is only an auxiliary variable. For the future situation however, it is of primary concern that VAT-data is successfully edited and not as much discarded. The following procedures are proposed in order to achieve this goal.

24. Aelen *et al.* (2005) have investigated the VAT-data of statistical units that have a covering percentage or filling percentage less than 100%. It turns out that, although these records are considered incomplete, VAT-turnover does match turnover from survey data as well as it does for complete enterprises. Also, analysis of historical VAT-data shows that the covering and filling percentage usually is a constant value in time. It is therefore likely that missing VAT-declarations for these fiscal units are only missing in theory, due to specific allocation of tasks of fiscal units, and that the resulting VAT-turnover is topical and complete.

25. Instead of using (*t*-1) as reference data, it might be better to use (*t*-*T*) (*T* being a year), since the Short-Term Statistics describe year-to-year developments. However, there are some disadvantages to that approach. An enterprise might experience a considerable development in one year, and consequently

could wrongfully be considered an outlier. Furthermore, chances are even higher that VAT-turnover is unknown at (*t-T).* Also, seasonal effects might shift from year to year. For instance, the Construction Industry Holiday, which traditionally has an important impact on the turnover in the Hotel and Catering Industry, shifts from year to year. By using a time series model (which would have to be able to handle missing values), above mentioned problems could be overcome, which would lead to a more realistic estimation of VAT-turnover at *t*.

26. As a basis for the outlier filter, a method is used that is originally developed to refine the editing of VAT-data as an auxiliary variable (Aelen *et al.*, 2005). The essence of this method, with some minor adjustments, is that $O^t_{BU,i}$ is an outlier when:

$$\frac{O^t_{BU,i}}{O'^t_{BU,i}} < \frac{1}{c_3} \text{ or } \frac{O^t_{BU,i}}{O'^t_{BU,i}} > c_4 \qquad (3)$$

$O^t_{BU,i}$ is the turnover and $O'^t_{BU,i}$ the reference value (described in the above sections) of enterprise *i* for period *t*. These values are also normalised with the median of the publication cell. Besides determining a realistic value for $O'^t_{BU,i}$, the crux of this method is to determine realistic parameters c_3 and c_4 (in the present editing procedure $c_3 = c_4 = 40$). Analysis of VAT-data on a number of lines of business has shown that the behaviour of fluctuations in VAT-turnover can be very specific for certain businesses. Also, fluctuations tend to be smaller for larger enterprises. Finally, fluctuations on VAT-turnover are generally larger for monthly declared turnover than for quarterly and yearly declared turnover. Therefore, it is suggested to determine c_3 and c_4 for different sub-populations, depending on the type of declaration (month, quarter, year), depending on a limiting turnover value for $O^t_{BU,i}$ or $O'^t_{BU,i}$ and depending on the business sector, SIC, or possibly even on the publication cell.

27. This method, which is quite uncomplicated in itself, requires a lot of effort to determine the above mentioned parameters. This is a problem that is very hard to tackle, since there is little information available on the quality of VAT-data, and it is not possible to confront enterprises with their VAT-declarations.

28. A possible solution is the use of VAT-assessments. As mentioned, these assessments can be considered to be a correction on the VAT-declarations. However, interpretation of these VAT-assessments is not straightforward. First of all, VAT-assessments are, like VAT-declarations, submitted per fiscal unit, and have to be converted to statistical units, with all the mentioned complications. Moreover, these assessments only contain the original declared sum and the corrected value, possibly including an imposed fine. The height of the fine depends on a number of factors, which include the possible history of negligence of the fiscal unit to pay or declare VAT, and the height of the assessment. Isolating the corrected assessment requires detailed information on the method of assigning penalties by the Dutch tax authorities, as well as historical data on the fiscal unit.

29. The declared VAT sum is a combined value, consisting of VAT on turnover with rates of 19%, 6% and 0%, depending on the sort of goods that are produced, decreased by VAT-return on purchased goods used for production. Typical lines of business will usually show similar combinations of production, thus resulting in a stable mean VAT-rate. Moreover, historical data of the concerning enterprise can be used for additional information. The difficulty lies with the VAT-return, since also paid VAT on incidental purchases (like investments in for instance machines) is settled in this value. However, the VAT-declaration contains detailed information of turnover for the various VAT-rates, and from that information the declared sum without VAT-return can be calculated.

30. Principally it is therefore possible to calculate the corrected VAT-turnover from VAT-assessments. With this information, the necessary parameters c_3 and c_4 can be calculated from historical VAT-data.

31. Depending on the number of outliers that are found with this procedure, a plausibility check could be performed to select a small number of outliers that are to be edited manually, while the bulk of the outliers will be edited automatically.

C. Macro-editing on the level of aggregates

32. When the edited VAT-data is aggregated, in a third step an editing procedure is proposed on the formed publication cells. VAT-turnover is compared with the corresponding aggregate for year (T–1). Score functions (Hoogland and Verburg, 2006) are applied to detect enterprises which cause inconsistencies between aggregates. These inconsistent turnover values can again be edited. This method is discussed in more detail in Van Haren *et al.* (2007) and Hoogland and Van Haren (2007).

V. CONCLUSIONS

33. Using VAT-data for Short-Term Statistics is certainly not straightforward. First of all, a number of problems need to be overcome in de process of converting VAT-data on the level of fiscal units into statistical data that can be used for publication.

34. Editing VAT-data is also quite a challenge. The main problem is that very little additional information is available on VAT-data. Besides historical information, knowledge on aggregates can be used. An extra source of information is the use of VAT-assessments, in order to calculate the necessary parameters that are used to determine outliers. Again, this procedure is quite complicated, but not impossible. The next step will be to actually determine these parameters for all assigned sub-populations.

References

Groen, Marja, Jeffrey Hoogland, Jos Jacobs (2007*), Methodology of converting VAT-data using the Source Data Base (SDB)* (in Dutch). Internal Unpublished research paper, Statistics Netherlands, Voorburg.

Hoogland, Jeffrey, Grietje van Haren (2007), *Editing and integrating VAT and SBS-data.* Unpublished research paper, Statistics Netherlands, Voorburg.

Loo, Eugène van (2007), *Decision tree on error detection Baseline/Source Data Base* (in Dutch). Internal Unpublished research paper, Statistics Netherlands, Heerlen.

Velzen, Jeroen van, Frank van de Pol, ed. (2006), *Handbook Short-Term Statistics, version 18* (in Dutch). Internal Unpublished research paper, Statistics Netherlands, Heerlen.

Aelen, Frank, Jan van den Brakel, Jeroen Ouwehand (2005), *Methodology Short-Term Statistics in Waves* (in Dutch). Internal Unpublished research paper, BPA-number TMO-R&D-2005-11-07-FALN, Statistics Netherlands, Heerlen.

Hoogland, Jeffrey and Ilona Verburg (2006), *Handling Inconsistencies in Integrated Business Data.* Invited paper for UNECE Work Session on Statistical Data Editing, 25-27 September 2006, Bonn, Germany.

Haren, Grietje van, Jeffrey Hoogland, Marette Kroonenberg, Ilona Verburg (2007), *ESB-integration phase 1: Approach and Results* (in Dutch). Internal Unpublished research paper, Statistics Netherlands, Voorburg.